Procesamiento del Lenguaje Natural, núm. 43 (2009), pp. 367-368 recibido 1-05-2009; aceptado 5-06-2009

CoCo, a web interface for corpora compilation*

CoCo, una interfaz web para la compilación de corpus lingüísticos

C. España-Bonet(1), M. Vila(2), H. Rodríguez(1), M.A. Martí(2)
(1) TALP Research Center
LSI Department
Universitat Politècnica de Catalunya
Jordi Girona 1–3, 08034 Barcelona
(2) CLiC
Linguistics Department
Universitat de Barcelona
Gran Via 585, 08007 Barcelona
cristinae@lsi.upc.edu, marta.vila@ub.edu, horacio@lsi.upc.es, amarti@ub.edu

Resumen: CoCo es una interfaz web colaborativa para la compilación de recursos lingüísticos. En esta demo se presenta una de sus posibles aplicaciones: la obtención de paráfrasis.
Palabras clave: Paráfrasis, Web Colaborativa, Interfaces

Abstract: CoCo is a collaborative web interface for the compilation of linguistic resources. In this demo we are presenting one of its possible applications: paraphrase acquisition.
Keywords: Paraphrasing, Collaborative Web, Interfaces

1. Introduction

CoCo[1] (Corpora Compilation) is a web interface designed for the compilation of linguistic corpora. Similar tools for a specific task or corpus can be found, such as the work by Chklovski (Chklovski, 2005b; Chklovski, 2005a) for collecting paraphrases or the Anawiki web page[2] by Poesio *et al.* (Poesio, Kruschwitz, and Jon, 2008) devoted to creating anaphorically annotated resources. As CoCo, these tools take advantage of web cooperation.

The system is open to any volunteer interested in contributing to the creation and widening of linguistic corpora, and it is currently being used by undergraduates at the University of Barcelona. CoCo will deal with different tasks, Paraphrasing, Coreference or Textual Entailment among them. The system is now prepared to gather data in four working languages: Catalan, Spanish, English and Arabic.

As stated previously, anyone can register and contribute as a user. Moreover, there is a subgroup of expert users which are allowed to control, modify and validate what is being incorporated into the database.

In the following section, we describe the task for which CoCo is currently being used: paraphrase acquisition.

2. Paraphrase Acquisition

Up to now, the operative part of the web is devoted to compiling a corpus of paraphrases. Paraphrases are understood as the different ways in which the same (or similar) content is expressed linguistically. There are two different approaches to the task. The first one, *General Paraphrasing*, aims to collect paraphrases of any kind. As a first input, the database has been filled with the paraphrases from the Microsoft corpus (Microsoft, 2005). The second, *Relational Paraphrasing*, is restricted to the paraphrases that express some kind of relationship between two entities. For now, it is devoted to the relationship of authorship.

2.1. General Paraphrasing

In the *Paraphrasing* section, the user is encouraged to widen the paraphrase corpus in three different ways.

- *Pair Generation.* A pair of paraphrasing sentences must be introduced.

* This research has been funded by the Spanish Ministry of Education and Science, project OpenMT (TIN2006-15307-C03-02), TEXT-MESS Lang2World (TIN2006-15265-C06/06), Ancora-Nom (FFI2008-02691-E/FILO) and the DOI/REFLEX-NBCHC050031 program as well as the FI Grant (2009FI_B 00690) from the Generalitat de Catalunya.

[1] http://www.lsi.upc.edu/∼textmess/

[2] http://www.anawiki.com/

- *Pair Completion.* Given a fixed original sentence, the user proposes a paraphrase. The original sentence is chosen from the existing corpora either randomly, sequentially or filtering by some criteria such as length or words contained.
- *Template Generation.* The same as in the previous task, but now the user is given part of the requested paraphrase. Some of the words are hidden so that the sentence only needs to be completed. The amount of hidden information can be modified by the user, who can hide or reveal words.

These three main tasks are accompanied by a section that allows users to search within the corpora or to modify their items.

Moreover, users subscribed as experts can evaluate already existing paraphrases. A pair is not accepted (or rejected) as a paraphrasing pair until it has been validated by at least three expert users.

2.2. Relational Paraphrasing

The second approach to Paraphrase Acquisition in the CoCo tool is that devoted to the collection of relational paraphrases. For now, the task focuses on *Authorship Paraphrasing*, that is, on those paraphrases that express some kind of relationship between an author and their work. We understand the relationship of authorship in a broad sense. It includes the relationship between painters and their paintings, between scientists and their theories, or between businessmen and their companies, to mention some examples.

As in the case of *General Paraphrasing* the user can choose different subtasks:

- *Authorship Generation.* A pair (author, work) is randomly shown and the user is asked to write a sentence containing the two items in an order which is randomly determined. A visual example of this task can be seen in Figure 1.
- *Web Evaluation.* Sentences automatically extracted from the web can be evaluated.

This section is already being used by students of Linguistics and Documentation at the University of Barcelona.

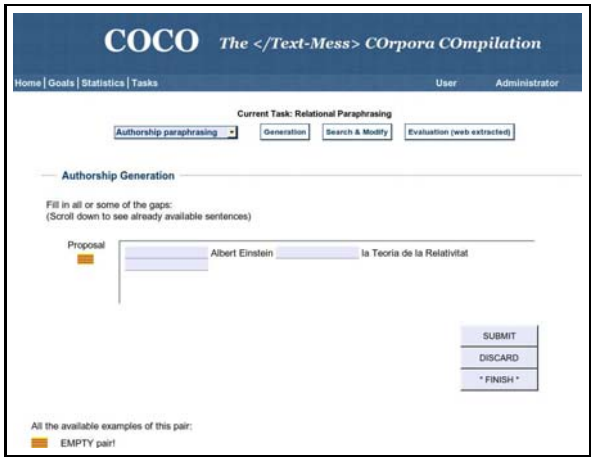

Figura 1: Screenshot of CoCo. This page allows the user to complete an authorship paraphrasing.

3. Conclusions

In this demo we are presenting a new tool for corpora compilation, currently being used for paraphrase acquisition. Up to now, the results obtained demonstrate CoCo's usefulness for the collection of corpora oriented to specific purposes. This is the case of the authorship paraphrase corpus that have been obtained for both Catalan and Spanish. The paraphrases obtained are being exploited in a current research on paraphrasing in the field of Linguistics.

References

Chklovski, Timothy. 2005a. 1001 paraphrases: Incenting responsible contributions in collecting paraphrases from volunteers. In *Proceedings of KCVC 2005*, pages 16–20.

Chklovski, Timothy. 2005b. Collecting paraphrase corpora from volunteer contributors. In *Proceedings of K-CAP 2005*, pages 115–120. ACM.

Microsoft. 2005. Microsoft research paraphrase corpus. http://research. microsoft.com/research/downloads.

Poesio, Massimo, Udo Kruschwitz, and Chamberlain Jon. 2008. Anawiki: Creating anaphorically annotated resources through web cooperation. In *Proceedings of LREC 2008*, pages 2352–2355.